Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars

Kuon, Joel-E ; Qi, Weihong ; Schläpfer, Pascal ; Hirsch-Hoffmann, Matthias ; von Bieberstein, Philipp Rogalla ; Patrignani, Andrea ; Poveda, Lucy ; Grob, Stefan ; Keller, Miyako ; Shimizu-Inatsugi, Rie ; Grossniklaus, Ueli ; Vanderschuren, Hervé ; Gruissem, Wilhelm

Abstract: Background Cassava is an important food crop in tropical and sub-tropical regions worldwide. In Africa, cassava production is widely affected by cassava mosaic disease (CMD), which is caused by the African cassava mosaic geminivirus that is transmitted by whiteflies. Cassava breeders often use a single locus, CMD2, for introducing CMD resistance into susceptible cultivars. The CMD2 locus has been genetically mapped to a 10-Mbp region, but its organization and genes as well as their functions are unknown. Results We report haplotype-resolved de novo assemblies and annotations of the genomes for the African cassava cultivar TME (tropical Manihot esculenta), which is the origin of CMD2, and the CMD-susceptible cultivar 60444. The assemblies provide phased haplotype information for over 80% of the genomes. Haplotype comparison identified novel features previously hidden in collapsed and fragmented cassava genomes, including thousands of allelic variants, inter-haplotype diversity in coding regions, and patterns of diversification through allele-specific expression. Reconstruction of the CMD2 locus revealed a highly complex region with nearly identical gene sets but limited microsynteny between the two cultivars. Conclusions The genome maps of the CMD2 locus in both 60444 and TME3, together with the newly annotated genes, will help the identification of the causal genetic basis of CMD2 resistance to geminiviruses. Our de novo cassava genome assemblies will also facilitate genetic mapping approaches to narrow the large CMD2 region to a few candidate genes for better informed strategies to develop robust geminivirus resistance in susceptible cassava cultivars.

DOI: https://doi.org/10.1186/s12915-019-0697-6

niklaus, Ueli; Vanderschuren, Hervé; Gruissem, Wilhelm (2019). Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. BMC Biology, 17(1):75.
DOI: https://doi.org/10.1186/s12915-019-0697-6

Kuon et al. BMC Biology (2019) 17:75
https://doi.org/10.1186/s12915-019-0697-6

BMC Biology

RESEARCH ARTICLE Open Access

Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars

Joel-E. Kuon[1*†], Weihong Qi[2†], Pascal Schläpfer[1], Matthias Hirsch-Hoffmann[1], Philipp Rogalla von Bieberstein[1], Andrea Patrignani[2], Lucy Poveda[2], Stefan Grob[3], Miyako Keller[1], Rie Shimizu-Inatsugi[4], Ueli Grossniklaus[3], Hervé Vanderschuren[5] and Wilhelm Gruissem[1,6*]

Abstract

Background: Cassava is an important food crop in tropical and sub-tropical regions worldwide. In Africa, cassava production is widely affected by cassava mosaic disease (CMD), which is caused by the African cassava mosaic geminivirus that is transmitted by whiteflies. Cassava breeders often use a single locus, *CMD2*, for introducing CMD resistance into susceptible cultivars. The *CMD2* locus has been genetically mapped to a 10-Mbp region, but its organization and genes as well as their functions are unknown.

Results: We report haplotype-resolved de novo assemblies and annotations of the genomes for the African cassava cultivar TME (tropical *Manihot esculenta*), which is the origin of *CMD2*, and the CMD-susceptible cultivar 60444. The assemblies provide phased haplotype information for over 80% of the genomes. Haplotype comparison identified novel features previously hidden in collapsed and fragmented cassava genomes, including thousands of allelic variants, inter-haplotype diversity in coding regions, and patterns of diversification through allele-specific expression. Reconstruction of the *CMD2* locus revealed a highly complex region with nearly identical gene sets but limited microsynteny between the two cultivars.

Conclusions: The genome maps of the *CMD2* locus in both 60444 and TME3, together with the newly annotated genes, will help the identification of the causal genetic basis of *CMD2* resistance to geminiviruses. Our de novo cassava genome assemblies will also facilitate genetic mapping approaches to narrow the large *CMD2* region to a few candidate genes for better informed strategies to develop robust geminivirus resistance in susceptible cassava cultivars.

Keywords: Cassava genomes, Cassava mosaic disease, Haplotigs, Optical mapping, Chromosome proximity ligation, Transposable elements, Allelic expression

Background

As a subsistence crop, cassava is valued for its starchy storage roots, especially by small-holder farmers, because the plant produces starch even under unfavorable environmental conditions. Cassava is also becoming increasingly important as an industrial crop and as livestock feed [1, 2]. But genetic gains from breeding in cassava have made little progress over the last century compared to other crops [3]. The heterozygous genome, long breeding cycles, clonal propagation, and poor asynchronous male and female flowering have limited substantial genetic improvement [4].

In Africa and India, cassava mosaic disease (CMD) is the most important economic threat for cassava production. The whitefly-transmitted virus is spreading and affecting agricultural productivity as a result of substantial yield losses in CMD-susceptible cultivars, in extreme cases up to 100% [5, 6]. An estimated 25 million tons of cassava storage roots are lost to CMD annually, impacting food security for more than 500 million people [7–9].

To date, only four geminivirus resistance genes (R-genes) have been identified, mapped, cloned, and characterized in crops [10–13], indicating that only a small

* Correspondence: joel-elias.kuon@biol.ethz.ch; wilhelm_gruissem@ethz.ch
†Joel-E. Kuon and Weihong Qi contributed equally to this work.
[1]Department of Biology, Institute of Molecular Plant Biology, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland
Full list of author information is available at the end of the article

proportion of the natural genetic diversity for geminivirus disease resistance has been exploited. For cassava, only three known genetic resistance loci present in the germplasm are currently providing relatively stable field resistance to CMD. These are the polygenic, recessive *CMD1* locus that was introgressed from wild cassava relatives [14], the single-dominant gene locus *CMD2* in tropical *Manihot esculenta* (TME) cultivars that confers resistance to all known CMVs [15, 16], and the resistance source *CMD3* that was distinguished from *CMD2* recently based on a single marker [17].

Because a single-dominant gene greatly facilitates breeding, the *CMD2* locus became the predominant resistance source deployed in African cassava breeding programs, although its underlying molecular mechanism and robustness are currently unknown. *CMD2* was discovered in landraces collected from farmer fields in Nigeria and other West African countries during the 1980s and 1990s, but the breeding pedigrees of these landraces are unknown [15]. Recently, the breakdown of the *CMD2* resistance during tissue culture-induced embryogenesis, which is an essential step in cassava transformation, was reported for TME cultivars [18]. The fact that many geminivirus resistance breeding programs rely on the stability of the *CMD2* locus makes it urgent to understand its genome organization and function. This can be achieved using high-quality de novo genome sequences for African cassava cultivars to fully exploit the importance of this resistance source.

Efficient crop plant genome sequencing is often constrained by genome size and heterozygosity as well as the excessive proportion of repetitive DNA elements (RE). The cassava genome has a haploid genome size of approximately 750 Mb [19], but its heterozygosity is among the highest found in sequenced plant genomes [20] and it is rich in REs. Thus, cassava genomes have proven difficult to assemble and to date only highly fragmented and incomplete genome assemblies are available [19–21]. The first cassava draft genome from the partly inbred South American genotype AM560 [21] was released in 2012, followed by draft genomes of an Asian cassava cultivar KU50 and the cassava wild relative W14 (*Manihot esculenta* ssp. *flabellifolia*) [20]. These genetic resources enabled first population genomic studies [16, 22–24], transcriptome characterization [25–27], and whole methylome profiling [28]. However, the current versions of the draft cassava genomes are represented as linear, haploid DNA sequences. Such a representation for highly heterozygous genomes can cause misleading results when using read mapping-sensitive applications that rely on accurate read placement [29]. For example, whole-transcriptome sequencing reads can align falsely or even fail to map when they span challenging regions with structural variations (SVs). Misplaced reads do in turn result in both missed true variants or incorrectly reported false variants and bias subsequent results.

Here we report the long read-based de novo assembled genomes of CMD-susceptible and *CMD2*-resistant African cassava cultivars as diploid-nature, haplotype-resolved chromosome assemblies. They were generated using single-molecule, real-time sequencing (SMRT; Pacific BioSciences) to assemble long haplotypes that cover multiple heterozygous regions. The continuity of the long-read genome assemblies was subsequently improved by contig scaffolding using long-range linking information from optical maps (BioNano) [30] and chromosomal conformation capture (Hi-C) [31, 32]. Furthermore, we generated full-length mRNA sequencing (Iso-Seq) to correct and improve predicted gene models. The two African cassava genome assemblies will facilitate the development of new heterozygous, haplotype-phased cassava reference-ready genomes and serve as a resource for the identification of causal *CMD2* resistance genes.

Results and discussion

Cassava genome sequencing, assembly, and chromosome-scale scaffolding

We achieved a nearly complete de novo diploid assembly and annotation of the genomes for the African cassava cultivars 60444, which is CMD susceptible, and TME3 that carries the dominant *CMD2* resistance (Fig. 1). Using 70× PacBio whole genome shotgun long reads with N50 read length of 12,813 bp (60444) and 12,424 bp (TME3), we assembled the TME3 genome into 12,971 contigs with a N50 of 98 kb (i.e., 50% of the assembly consists of 98 kb or longer contigs). The 60444 genome was assembled into 11,459 contigs with a N50 of 117 kb (Table 1) (Additional file 1: Figure S1, Additional file 2: Table S1). We evaluated the performance of three different long-read assemblers (CANU-MHAP [34], FALCON v0.5 [35] and PBcR-MHAP [36]) by aligning Illumina paired-end (PE) reads to the corresponding long-read assemblies. This showed that the CANU assembler generated the most accurate assemblies, with the highest proportion of mapped paired-end (PE) reads (98.4% for 60444 and 96.4% for TME3) and the lowest proportion of discordant read-pair alignments (1.6% for TME3 and 1.2% for 60444) (Additional file 2: Table S2).

The total length of assembled contigs was above 900 Mb for both TME3 and 60444. This was higher than the haploid genome size of approximately 750 Mb estimated by flow cytometry (Additional file 1: Figure S2), indicating that haplotypes of the heterozygous genomes were assembled independently into different contigs [37, 38]. Based on contig alignments against each other and read depth of coverage, we reassigned allelic contigs as primary contigs and haplotigs using Purge Haplotigs [39]. The total size of the de-duplicated primary haploid

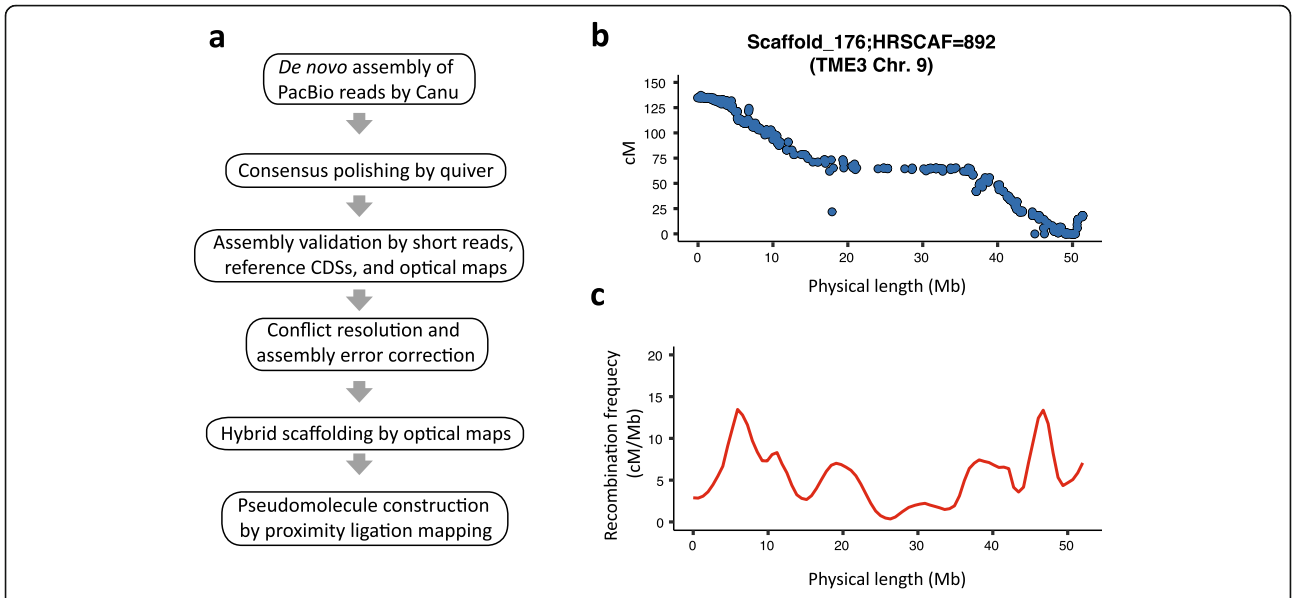

Fig. 1 Assembly pipeline for the 60444 and TME3 African cassava genomes. **a** Overview of the processing pipeline used for the assemblies (see Additional file 3 for details). **b** Pseudomolecule validation using the location of SNP markers on the physical map (*x*-axis) as compared to their position on the composite cassava genetic map [33] (*y*-axis) for Chr.9, which is a single scaffold 176_{TME3}. Each genetic marker is depicted as a dot on the plot (937 data points). **c** Graphical representation of mean local recombination frequencies between SNP markers along scaffold 176_{TME3}. The *x*-axis represents the physical positions of the means on Chr 9, and the *y*-axis indicates the recombination ratio (centiMorgan (cM)/Mb) in each 1-Mb sliding window

assembly was 732 Mb for TME3 and 713 Mb for 60444 (Table 1), which was close to the flow cytometry measurement (Additional file 1: Figure S2). The secondary haplotig assembly was more than 200 Mb. This reflects the high heterozygosity within the cassava genome, which is the consequence of interspecific admixture and past breeding, but short runs of homozygosity are also present in the genome [19, 40]. In this case, optical mapping is useful to phase haplotypes, especially in genomes with divergent homologous chromosomes [41]. We generated two high-coverage optical maps (150× for 60444, 130× for TME3) using the BioNano Genomics IrysView DNA imaging and analysis platform. The fluorescently labeled DNA molecules of the two cassava genomes assembled into similarly sized genomes of 1205 Mb for TME3 and 1204 Mb for 60444. This indicates that most of the parental chromosomes had been "phased" into haplotype segments by optical mapping (Additional file 2: Table S3). To further improve sequence contiguity and haplotype phasing, the PacBio

Table 1 Assembly statistics for the cassava TME3 and 60444 genomes compared with previously published assemblies of cassava genomes

Cultivar	TME3	60444	KU50 [20]	AM560 [19]
Number contigs	12,971	11,459	99,509	39,574
Contig N50 (kb)	97.58	116.8	5.28	27.87
Total contig length (Mb)	947	975	NA	NA
Total primary contig length (Mb)	732	713		
Total haplotig length (Mb)	213	260		
Optical map supported scaffolds	558	552	NA	NA
Primary scaffolds	506	491		
Optical Hybrid-scaffold N50 (Mb)	2.25	2.35	NA	NA
Hi-C scaffolding N50 (Mb)	53.35	59.19	NA	NA
Assembly size (Mb)	1225	1277	291.1*	582.3
TE proportion (%)	64.81	64.91	25.7	50.3
Annotated protein-coding genes	33,853	34,127	38,845	33,033

*The KU50 genome was reported to be 495 Mb in [20]; the number shown here was the published and downloadable DNA sequence available in 2014

contigs were corrected, joined, ordered, and oriented according to the optical mapping data. This generated a set of 558 optical-map-supported scaffolds spanning 634.1 Mb with a scaffold N50 of 2.25 Mb for TME3. For 60444, we generated 552 scaffolds spanning 714.7 Mb with an even higher scaffold N50 of 2.35 Mb.

The Portuguese introduced cassava from South-America to Africa in the sixteenth and seventeenth century, and since then the African germplasm diversity has remained exceptionally narrow [42]. Previous diversity studies relied on short-read mapping data only, but genome-wide structural variants are challenging to detect in heterozygous and complex plant genomes. The diploid optical maps from the two African cassava cultivars were tested for genomic diversity. The vast majority (81%) of the consensus optical maps from TME3 could be aligned with those from 60444 via common label patterns, indicating a very low level of structural diversity between the two cassava genomes. We then screened the alignments for TME3-specific insertions and deletions (INDELs) and identified evidence for 1058 insertions and 1021 deletions with average sizes of 57.4 kb and 45.7 kb, respectively (Additional file 2: Table S4).

Genome completeness and haplotype phasing

Haplotype phasing, or identifying alleles that belong to the same chromosome, is a fundamental problem in genetics. Our assembly strategy using PacBio long reads in combination with BioNano optical maps produced haplotype-aware genomic scaffolds in which phase information over long regions of homozygosity and even across assembly gaps was resolved. To further assess the completeness and quality of phased haplotypes in the two cassava genomes, publicly available cassava coding DNA sequences (CDSs) [19] were aligned to each of the assembled optical scaffolds using GMAP [43], which takes into account exon-intron junctions. Local duplicates, i.e., inter-scaffold matches, and CDSs with < 99% alignment coverage were removed from the analysis. Of the 41,381 CDS, 99.93% are present in the 60444 and TME3 genomes with only a few missing (84 and 86, respectively). This CDS alignment was used to estimate the haplotype phasing and allele number variation. In total, we detected 18,831 and 19,501 multi-copy gene loci in TME3 and 60444, respectively, with a large proportion of CDS aligning into allelic pairs (n = 15,679 for TME3 and n = 17,019 for 60444) (Fig. 2a).

Centuries of cassava clonal propagation has resulted in genetically fixed deleterious mutations that affect crop vigor and strongly limit breeding [3, 44, 45]. Duplicated regions are often subject to dynamic changes, including the accumulation of point mutations that facilitate species diversification [46]. To test this hypothesis for the bi-allelic genes in the diploid 60444 and TME3 genomes, we measured the nucleotide diversity for each allelic pair as determined by AM560 CDS alignments and plotted the proportion of single-base pair mutations. This analysis revealed high variation between coding sequences of alleles, further substantiating the heterozygosity within the coding portion of the genome (Fig. 2b) (sequence alignment mean: allele I-A 99.26%, allele I-B 97.15%).

Short-read-based genome assemblies frequently do not capture intergenic sequences that might be important for gene regulation because promoter regions often adjoin repetitive DNA sequences. Investigating gene expression regulation is highly dependent on accurately assembled promoters. We screened the promoter regions of bi-allelic genes and analyzed their sequence contiguity over a 20-kb region upstream the translational start codon (Fig. 2c). This revealed near complete promoter regions in the 60444 and TME3 genomes as compared to the AM560 v6.1 genome. The extensive sequence contiguity will facilitate allele-specific expression analysis and the identification of novel tissue-specific cassava promoter sequences.

To determine if the accumulation of allelic mutations has an impact on gene expression, we measured allele-specific expression using high-throughput RNA-seq analysis from eight sequencing libraries that originated from different tissues (for details, see Additional file 3). In total, we covered the expression of 18,723 genes with two alleles and identified 3451 (14.43%) genes with mono-allelic expression (Fig. 2d, e). Various mono-allelic expressed genes (44.76%) have highly similar promoter sequences (mean similarity = 95.52%) between the alleles, indicating that mono-allelic expression of these genes could be caused by one or more SNPs or may be epigenetically regulated through DNA methylation or chromatin packaging. It has been suggested that cassava developed a more robust maintenance methylation mechanism than found in other crop plant species [28]. The high number of alleles not expressed in the analyzed tissues could be another property of the cassava genome that was maintained through clonal propagation of the crop over centuries.

Assembling pseudochromosomes of heterozygous cassava genomes

In cassava, a single bi-parental cross rarely yields enough progeny to generate a robust and dense genetic map that can be used to genetically anchor sequences to chromosomal pseudomolecules. The most recent publicly available cassava composite genetic map was generated from various mapping populations and anchors only 71.9% of an earlier haploid genome assembly [33]. To re-construct the set of cassava chromosomes independently of a composite genetic map (i.e., de novo), we generated chromosome proximity ligation libraries (Hi-C) for the TME3 and 60444 cassava cultivars (for details, see Additional file 3).

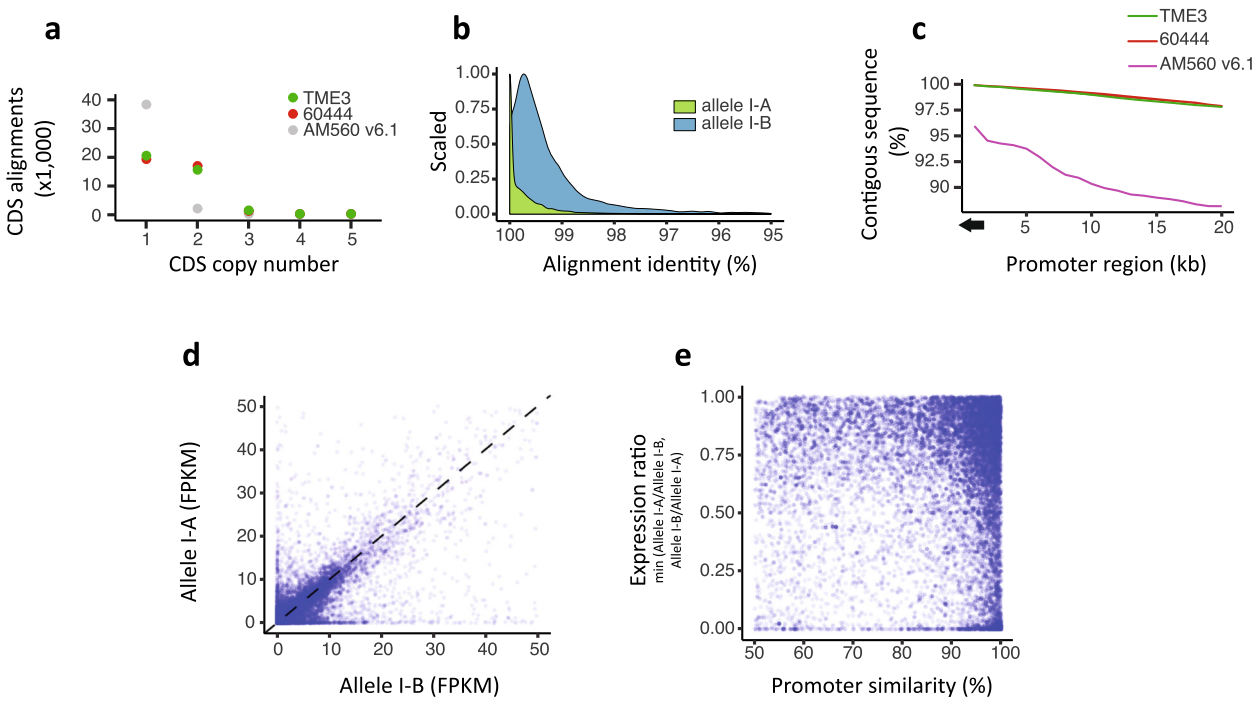

Fig. 2 Haplotype phasing, allele nucleotide diversity, and allele-specific expression analysis for haplotype-aware cassava genomes. **a** Cassava CDS collection ($n = 41,381$) and their alignment copy number distribution in the two African cassava genomes TME3 (green points) and 60444 (red points), and the AM560 v6.1 genome (gray points). **b** Sequence alignment properties for the bi-allelic reference CDSs ($n = 13,425$) found in the 60444 genome. Bi-allelic genes, depicted as allele I-A and allele I-B, are presented as a green curve, and the homologous allelic counterpart as a blue curve. Percentage of alignment identity is shown on *x*-axis and data point density on the *y*-axis. **c** Promoter sequence contiguity ("N"-free-sequence) comparison between three different cassava genomes measured using 1-kb bins over a 20-kb region upstream the transcriptional start site. **d** Scatterplot of allele-specific gene expression in 60444 based on RNA read counts measured as fragments per kilobase of sequence per million mapped reads (FPKM). A bi-allelic gene is depicted as a single blue dot. Expression of one allelic copy is shown on the *x*-axis and the expression of the homologous counterpart on the *y*-axis. **e** Bi-allelic gene expression as a function of promoter sequence similarity. The bi-allelic gene expression ratio (*y*-axis) of 1.00 indicates an equal expression of both alleles, whereas the expression ratio of < 0.25 indicates mono-allelic expression ($n = 3451$). The promoter sequence similarity between the homologous alleles measured for a 2-kb region upstream of the start codon is shown on the *x*-axis. Bi-allelic genes with identical or near-identical promoter sequences can have mono-allelic expression

Proximity mapping was previously shown to be instrumental for chromosome-scale assemblies in other species [31, 32]. The optical-map-improved scaffolds were combined with the remaining contigs and grouped according to the Hi-C-based molecule interaction maps using Dovetail proprietary algorithms. The approach has already been used recently in other crop genome sequencing projects to generate pseudochromosomes from the assembly of contigs and smaller scaffolds into contiguous scaffolds of chromosome size [47, 48]. Implementing the Dovetail assembly for cassava increased sequence contiguity by nearly 25-fold for a final scaffold N50 of 53.4 Mb in the TME3 and 59.2 Mb in the 60444 in African cassava genomes.

To assess the quality of the Hi-C-based chromosomal pseudomolecules, we aligned the genetic markers from the cassava composite genetic map [33]. Out of 22,403 genetic markers, we were able to align 22,341 (99.7%) with the 60444 genome and 22,373 (99.8%) with the TME3 genome. To visualize and validate the chromosomal pseudomolecules, we plotted the genetic distance against the physical distance for each genetic marker. At this level of resolution, these plots confirm that whole pseudochromosomes were assembled without large inter-chromosomal re-arrangements (Fig. 1b, Additional file 1: Figure S4). Plotting the recombination rate using a sliding window of 1 Mb across assembled scaffolds revealed the expected decrease in recombination frequency in the center of the scaffold, as well as the presence of other regions with low recombination in the chromosome arms (Fig. 1c, Additional file 1: Figure S5).

When analyzing the fasta sequences of the cassava pseudochromosomes in more detail, we found TME3 and 60444 pseudochromosomal scaffolds to contain more DNA sequence compared to the AM560 genome (Additional file 1: Figure S6). For example, Scaffold 7_{TME3} and Scaffold 1478_{60444} representing chromosome 12 were 107.1% and 116.3% larger than the chromosome 12 in AM560. The total length of the TME3 and 60444 pseudochromosomes was 29% greater than the haploid genome size estimated by flow cytometry, respectively. The additional sequences

originate from repetitive sequences or spacers that were added by Dovetail in the assembly process but also represent coding sequences and gene models as well. When aligning the haploid composite genetic map [33] to the genome, we noticed that for loci where both haplotypes were assembled as allelic contigs/scaffolds, Hi-C scaffolding tended to integrate both haplotypes into pseudochromosomes, thus inflating genome size. We identified 78% of the genetic markers in TME3 (82.8% in 60444) as perfect hits (100% identity and coverage). Of those, 29.1% were present more than once in the TME3 genome (29.8% of 60444) (Additional file 1: Figure S7). Such a multiplication was expected, since both TME3 and 60444 are heterozygous genomes. We analyzed the various genome assemblies and found that the numbers of genetic markers that were present more than once were constant throughout the assembly process. In the CANU and CANU-BNG assemblies of both TME3 and 60444, the genetic markers are predominantly on different contigs and scaffolds, confirming that haplotypes have been assembled into separate allelic sequences. This is different in the Dovetail pseudochromosomes (Additional file 1: Figure S4), where 54.8% of TME3 and 56.5% of 60444 genetic markers can be found on contiguous sequences more than once (Additional file 1: Figure S7 E–F), indicating that both haplotpyes have been lifted up into Hi-C scaffolds. Co-location of genetic markers on the same scaffold was not a local phenomenon but was spread over the entire genome. For example, on scaffold 7_{TME3} representing presudochromosome 12 (Additional file 1: Figure S8), 2635 genetic markers are aligned twice or more, while they were mostly separated on allelic sequences in the CANU-BNG assemblies, indicating integration of both haplotypes in the Dovetail pseudochromosome (Additional file 4: Table S5). Copies of the same genetic marker typically occur in close proximity to each other, with a median distance of 343 kb. A remaining set of 87 genetic markers was already duplicated on individual contigs of scaffold 7_{TME3} in the initial CANU assembly of chromosome 12 and thus likely represent true gene duplication events. They were on average separated by 27.9 kb with up to eight gene copies per contig in some cases. After removing the duplicated allelic sequences in the Dovetail pseudochromosomes based on haplotig purging (Additional file 2: Tables S6 and S7), the total size of the pseudochromosomes was 796 Mb for TME3 and 854 Mb for 60444.

Proximity ligation mapping was also used to identify miss-joints and mis-assemblies. Based on the Hi-C data, we identified 30 mis-assemblies in the TME3 optical map scaffolds and only 16 in the 60444 scaffolds. Each mis-assembly was validated manually by testing Hi-C read-pair alignment positions and alignment depth, and scaffolds were corrected accordingly (Additional file 1: Figure S9). However, the proximity maps of TME3 and 60444 will be valuable for quality assessment of the composite genetic map and to improve the sequence resolution in regions that are seemingly devoid of meiotic recombination.

Repetitive DNA analysis and genome annotation of cassava pseudochromosomes

Transposable elements (TEs) and REs are involved in genome evolution and shaping gene regulatory networks [49]. Unlike previous sequencing technologies, SMRT reads can span and resolve entire TE and RE regions [50]. Using de novo generated cassava DNA repeat libraries, we annotated up to 2.5 times more TEs (defined by REPEATMASKER and REPEATMODELER, as described in the "Methods" section) in the pseudochromosomes compared to earlier reports [19–21] (Fig. 3a). In the TME3 and 60444 Dovetail assemblies, we annotated 602.90 Mb (64.81%) and 633.93 Mb (64.91%) as repetitive sequences, respectively. As an example, we investigated the spatial distribution of sequence repeats along the entire chromosomal scaffold 1583_{60444}, which corresponds to pseudochromosome 9 (Fig. 3b) and generated density maps for the four predominant TE categories. Long terminal repeat (LTR) retrotransposons have higher densities in the centromer region, while non-LTR retrotransposons elements (LINE and SINE) are clustered in telomere-proximal regions. Class II DNA transposons are more equally distributed across that scaffold. A similar distribution of TEs was reported for other complex plant chromosomes [51, 52], confirming the high quality of cassava genome sequences ordered using Hi-C. Our pseudochromosome assemblies reveal a high proportion of repetitive DNA in cassava (65% of total contig length), which is similar to the amount of repetitive DNA found in other sequenced complex crop genomes such as sorghum (54%) [53], quinoa (64%) [54], or barley (81%) [52] (detailed TE annotation in Additional file 2: Table S9).

We predicted protein coding and microRNA (Additional file 2: Table S10) sequences using a combination of ab initio prediction and transcript evidence from available cassava gene models [19]. Protein-coding sequence annotation was assisted by Iso-Seq (high-quality, full-length cDNAs from single-molecule sequencing) data that covered 15,478 (45.7%) gene loci in TME3 and 16,057 (47.0%) in 60444 (Additional file 1: Figure S10). The quality of the gene model annotation was assessed for 1440 conserved plant genes using BUSCO [55]. We found 95% of the single-copy conserved orthologs in both genomes, with only 20 and 19 genes partially assembled in TME3 and 60444, respectively (Additional file 2: Table S11).

Protein expansion in cassava genomes

The two African cassava cultivars 60444 and TME3 are thought to have exceptional low genetic diversity [19]. The similar number of annotated genes allowed us to investigate gene family expansions specific to the two

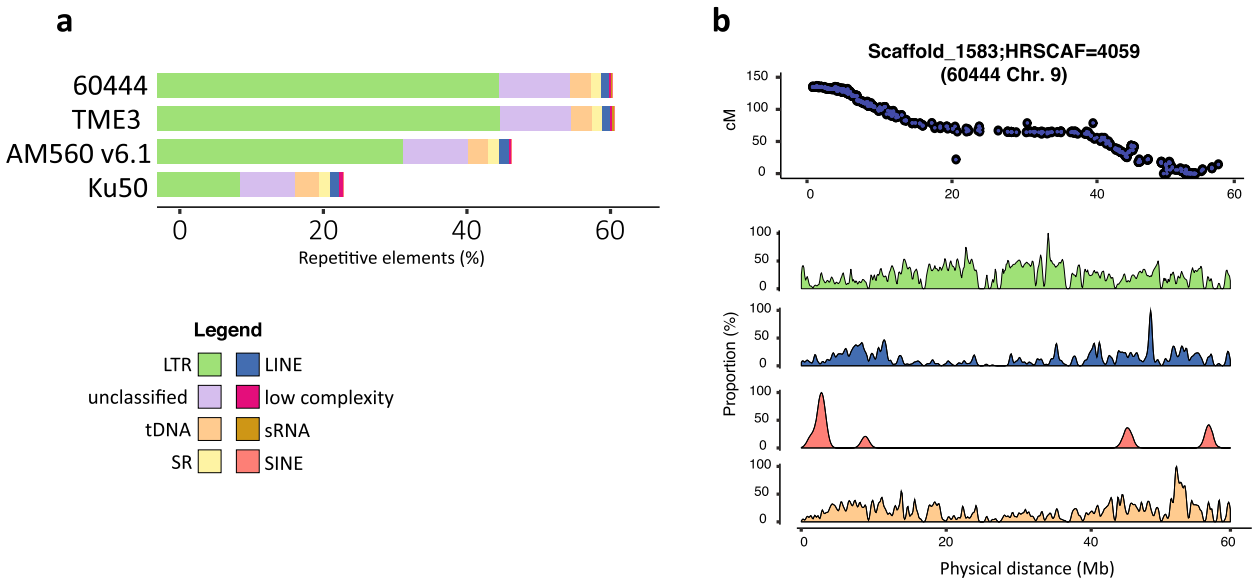

Fig. 3 Distribution of major repetitive elements identified in cassava genomes. **a** Percentage of base pairs of assembled cassava genomes that represent long terminal repeat (LTR), unclassified repeat, DNA transposon (tDNA), sources of short RNA (sRNA), long interspersed elements (LINE), low-complexity element, and short interspersed nuclear element (SINE) sequences. **b** Graphical representation of SNP markers (top) and chromosomal density plots for the four predominant TE categories (bottom) on the scaffold 1583_{60444} map

cultivars. We used OrthoMCL clustering of all gene models present in our two assemblies as well as the genome assemblies of the South American cassava cultivar AM560, *Ricinus communis* as a close relative of cassava, and *Arabidopsis thaliana* as an outgroup [56, 57]. This confirmed that the two African cassava cultivars are closely related (Fig. 4a). For example, there were fewer gene family groups specific to 60444 or TME3 (0.8–1.1%), whereas the number of specific gene family groups was considerably larger for *Ricinus* and *Arabidopsis*. Interestingly, there were more protein groups associated exclusively with AM560 and *Ricinus* than with *Ricinus* and either 60444 or TME3. These trends were also seen for predicted enzymatic reactions (Fig. 4b) and predicted metabolic pathways (Fig. 4c) but, as expected, overall the four species were similar for total reactions and metabolic pathways [57].

There remained 1823 protein groups containing 4081 gene models (2067 for 60444 and 2014 for TME3) that are specific to the two African cassava genomes. Considering the short evolutionary time since cassava was introduced to Africa about 400 years ago, it is likely that the differences in gene divergence and expansions between AM560, 60444, and TME3 evolved before the ancestor or ancestors of 60444 and TME3 was brought to the African continent.

We subsequently investigated genes of proteins associated with gene families for overrepresentation of GO terms [58]. For AM560, we found cultivar-specific proteins with GO terms enriched for "polygalacturonase activity" (Additional file 1: Figure S11). Among the most significantly enriched GO terms for genes that were associated exclusively with the African cultivars were categories "structural integrity of ribosomes" (GO:0003735) and "structural molecule activity" (GO:0005198) (Additional file 1: Figure S12). Another more specific function was squalene monooxygenase activity (GO:0004506). Interestingly, single-strand DNA virus infection increases squalene production [59]. Squalene monooxygenase converts squalene to (3S)-2,3-epoxy-2.3-dihydrosqualene (epoxysqualene), which is a precursor for many specialized metabolites (Additional file 1: Figure S13). Both in 60444 and TME3, there are four metabolic pathways predicted to be involved in the conversion of epoxysqualene to several specialized metabolites. Some have known antimicrobial, anti-inflammatory, and/or anti-tumor activities, including beta-amyrin that can be converted to oleanolate, which has antiviral activity [60] and inhibits topoisomerase I/II [61], which are involved in replication of viruses such as cauliflower mosaic virus (CaMV) [62] . The Rep locus in the CMD-related mungbean yellow mosaic virus (MYMV) encodes a protein with topoisomerase activity [63]. Since the Rep locus is found in all Gemini viruses, functionality is likely conserved [64]. The pathway from squalene to oleanolic acid involves three consecutive reactions that all have gene annotations in all three cassava cultivars. The two African cultivars 60444 and TME3 that are exposed to CMGs, however, have an expanded gene pool for two of the three reactions in the pathway (Additional file 1: Figure S12).

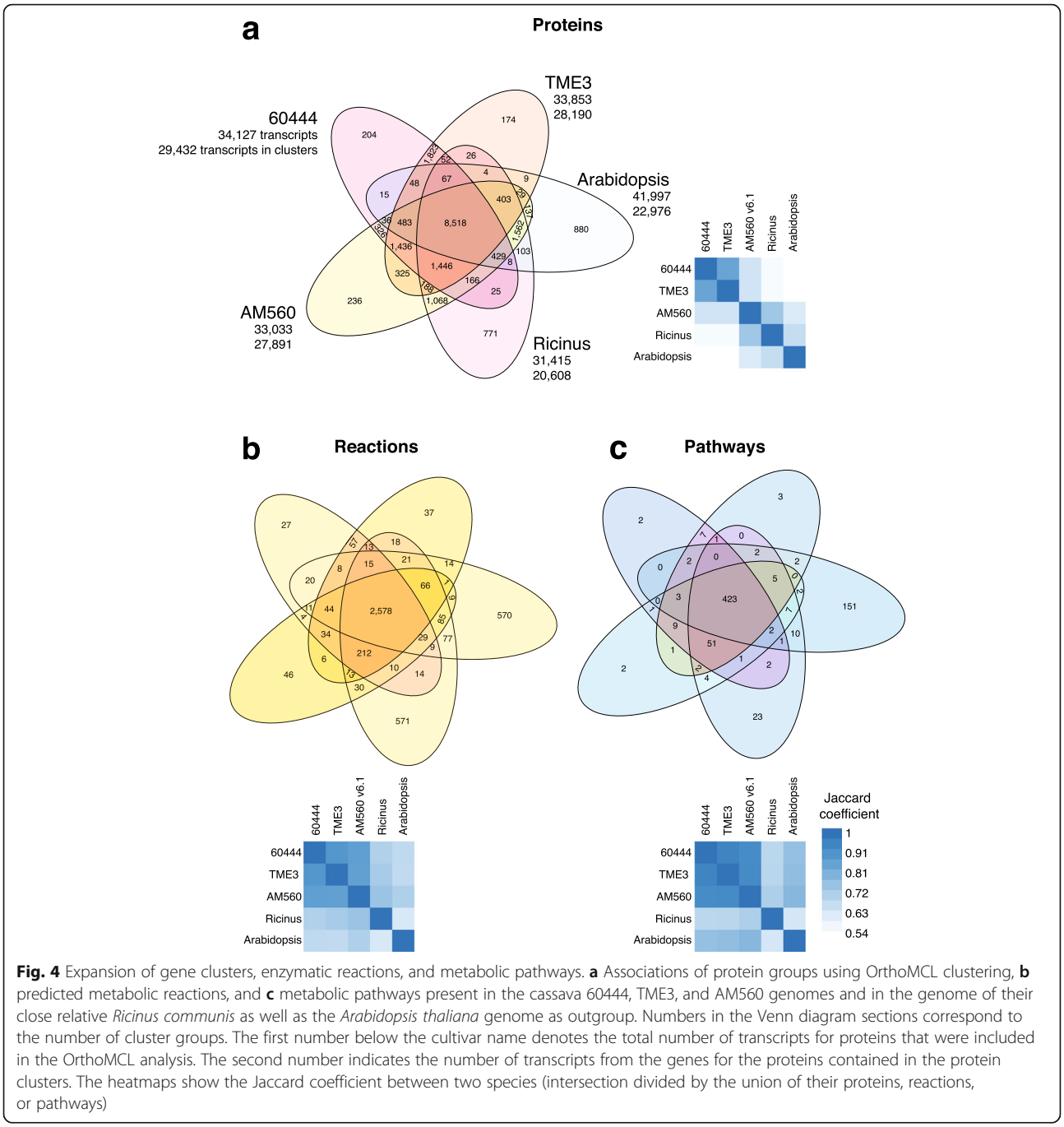

Fig. 4 Expansion of gene clusters, enzymatic reactions, and metabolic pathways. **a** Associations of protein groups using OrthoMCL clustering, **b** predicted metabolic reactions, and **c** metabolic pathways present in the cassava 60444, TME3, and AM560 genomes and in the genome of their close relative *Ricinus communis* as well as the *Arabidopsis thaliana* genome as outgroup. Numbers in the Venn diagram sections correspond to the number of cluster groups. The first number below the cultivar name denotes the total number of transcripts for proteins that were included in the OrthoMCL analysis. The second number indicates the number of transcripts from the genes for the proteins contained in the protein clusters. The heatmaps show the Jaccard coefficient between two species (intersection divided by the union of their proteins, reactions, or pathways)

CMD2 locus

The identification and molecular characterization of geminivirus resistance genes in cassava has been slowed by missing genomic resources. Previous genetic mapping placed the *CMD2* locus in separate regions of AM560-2 (v6.1) chromosome 12 [16, 22], suggesting that precise *CMD2* mapping is difficult because of few recombination events and borderline marker saturation. We found that genetic markers released from these mapping efforts aligned to an approximate 5-Mb region between 49 and 55 Mb of scaffold 7_{TME3} (Fig. 5a). The same markers were identified on 60444 scaffold 1478_{60444}.

Analysis of the *CMD2* locus in scaffold 7_{TME3} revealed that nearly all markers from a bi-parental mapping population [16] aligned to a region between 51 and 55 Mb (Fig. 5a, red circles, with a single marker outside of this region at 49 Mb) and the marker set that had been generated from an association mapping approach [22] spanned an adjacent region of approximately 3 Mb (49–51 Mb) in the same scaffold (Fig. 5a, blue circles). These

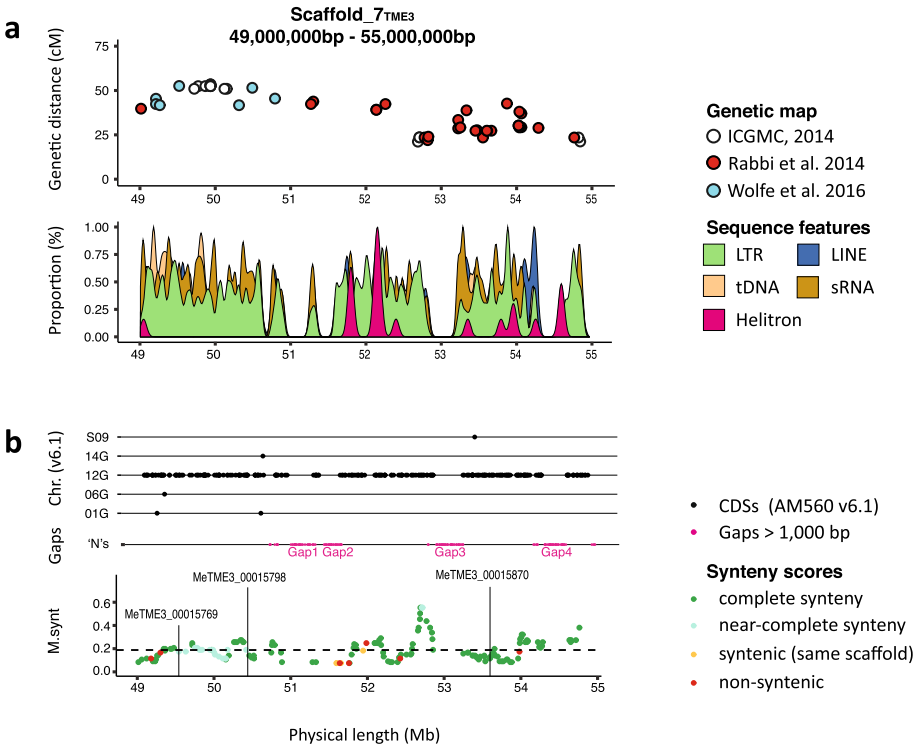

Fig. 5 *CMD2* locus in TME3 genome. **a** The upper panel shows *CMD2*-associated genetic SNP markers and their genetic distance relative to their physical position on scaffold_7 of TME3. Red dots indicate *CMD2* SNP markers released by Rabbi and colleagues [16], and blue dots indicate the SNP markers released by Wolfe and colleagues [22, 42]. The lower panel shows the distribution of main repetitive genomic features at the *CMD2* locus. **b** The upper panel shows the alignment position of AM560 v6.1 CDS in the region of Chr. 12 containing the *CMD2* locus. Each black dot represents the CDS alignment position at the *CMD2* scaffold (*x*-axis) and its chromosomal origin from the AM560 v6.1 cassava reference genome. Sequence breaks (gaps > 1 Kb) are shown as pink bars. The lower panel shows the MSS for every annotated gene at the *CMD2* locus in TME3. Green dots indicate genes that are found in the *CMD2* region of 60444, and light blue dots indicate genes that are found in close proximity of the *CMD2* locus in 60444. Orange dots indicate TME3 genes that show a syntenic relation to 60444 genes on other 60444 scaffolds, and red dots indicate genes with no syntenic relation. The dashed line represents the MSS average for the whole genome

results suggest that the genetic marker sets that previously identified two separate loci in fact correspond to a single region spanning 6 Mb of scaffold 7_{TME3}. However, the pseudochromosome 12 region containing the *CMD2* locus has four major assembly gaps (Fig. 5b), which likely result from extensive stretches of repetitive DNA that prevent complete assembly of the region. The alignment of the AM560 CDS in the *CMD2* region revealed high conformity with the AM560 chromosome 12 to scaffold 7 of TME3 containing the *CMD2* locus (Fig. 5b). In 60444, the markers aligned with a 6-Mb region on Scaffold 1478_{60444}.

To better understand the similarity between the 60444 and TME3 genomes, we analyzed their synteny and in particular synteny in the region of the *CMD2* locus using the Comparative Genomics platform (CoGe) (Additional file 1: Figure S14). More than 70% of the genes encoded within the $CMD2_{TME3}$ locus were found to be syntenic to a gene within the $CMD2_{60444}$ and $CMD2_{AM560}$ loci (Fig. 5b, Additional file 1: Figure S15). Less than 10% of the genes either had no syntenic gene (red) in the other two genomes or the syntenic genes were outside the *CMD2* locus in a larger region three times the size of the *CMD2* locus. Two TME3 genes, MeTME3_00015756 and MeTME3_00015762, are missing from the *CMD2* regions of AM560 and 60444, both short gene models of unknown functions. While at the level of microsynteny most genes are syntenic, the organization of the *CMD2* locus is not entirely contiguous between the TME3, 60444, and AM560 genomes except for a region with high microsynteny around 52.7 Mb. It is unlikely that the low organizational microsynteny is the result of pseudochromosome mis-assemblies because genes between 52.1 and 54.7 Mb of $CMD2_{TME3}$ are found on a single CANU-BNG scaffold with low microsynteny to the corresponding regions in AM560 and in 60444.

We searched our de novo gene annotations in the *CMD2* loci of the TME3 and 60444 chromosome 12 scaffolds for three suggested CMD resistance candidate genes that were identified in the AM560 v6.1 genome [22]. Manes.12G076200 and Manes.12G076300 encode peroxidases, a protein class that is involved in many biochemical reactions [65]. In tomato, peroxidase activity

increases in juvenile leaves during whitefly-mediated geminivirus infections [66]. We confirmed the presence of the two peroxidase genes (MeTME3_00015769 and MeTME3_00015798) at the *CMD2* locus of 60444 and TME3. Manes.12G068300 encodes a protein disulfide-isomerase-like 2-3 (PDI). This type of enzyme catalyzes the correct folding of proteins and prevents the aggregation of unfolded or partially folded precursors. We identified MeTME3_00015870 in the *CMD2* locus of TME3 that encodes a similar PDI. In barley, genetic studies identified HvPDI5-1, which is the ortholog of MeTME3_00015870, as a virus susceptibility factor that contributes to resistance to bymoviruses [67].

When expanding the search proximal and distal to the *CMD2* locus for genes that could provide resistance to geminivirus infection, we identified a gene encoding Suppressor of Gene Silencing 3 (SGS3, MeTME3_00015743, 1.71 Mb downstream of the *CMD2* locus). SGS3 is involved in posttranscriptional gene silencing (PTGS) and functions together with RNA-directed RNA polymerase 6 (RDR6) during dsRNA synthesis [68]. SGS3 has also been suggested to function in the transport of the RNA-silencing signal [69]. SISGS3, the tomato homolog of Arabidopsis SGS3, interacts with the tomato yellow leaf curl geminivirus (TYLCV) V2 protein that functions as a suppressor of silencing and counteracts the innate immune response of the host plant [70]. The identified genes provide useful information for candidate proteins related to the function of the dominant *CMD2* locus in protection against geminivirus infection in TME3 and other *CMD2*-type cassava cultivars.

Conclusions

The diploid-aware de novo assemblies of the heterozygous 60444 and TME3 cassava genomes will help to unlock the limited genomic diversity of African cassava cultivars for crop improvement and geminivirus resistance breeding. The genome assembly strategy reported here can be similarly adapted to other medium-sized, non-inbred genomes with high heterozygosity and DNA repeat-rich regions. Using the information for haplotype-phased alleles and allele-specific expression, it will be possible to characterize and purge deleterious mutations using targeted genome editing [71], conventional breeding, or genomic selection. Moreover, the large haplotype scaffolds of the 60444 and TME3 genomes will greatly facilitate trait mapping and map-based cloning of agriculturally important genes in this important food security crop.

Our results show that the new maps of the *CMD2* locus in both 60444 and TME3, together with the newly annotated genes, will help to identify the causal genetic basis of *CMD2* resistance to geminiviruses. Our de novo genome assemblies will also facilitate genetic mapping efforts to narrow the large *CMD2* region to a few candidate genes for better informed strategies to develop robust geminivirus resistance in susceptible cultivars. Furthermore, the genome assemblies will lead to a better understanding of the genetic differences between cassava cultivars and how genetic variability can be deployed in breeding programs for future cassava improvement.

Methods

Further details of all methods are presented in Additional file 3. No statistical methods were used to predetermine sample size. Experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Long-read sequencing and sequence assembly

To sequence the two cassava genomes with long reads, we extracted high molecular weight (HMW) genomic DNA from 3-week-old leaf tissue of in vitro grown cassava 60444 and TME3 plants following a modified protocol [72]. Libraries for PacBio SMRT sequencing were generated as described previously [73]. Libraries were sequenced using a PacBio RSII instrument with P6C4 sequencing reagents. We used 47 SMRT cells for TME3 and 45 SMRT cells for 60444. For 60444, we generated a total of 52.4 GB with subread bases with a mean read length of 12.8 kb. For TME3, 53.9 GB of subread bases were generated with a similar mean read length of 12.4 kb. The PacBio sequences had a > 70-fold genome coverage.

De novo assembly of the subreads was performed applying three assemblers: the PBcR-MHAP pipeline [36], the CANU-MHAP assembler [34], and the FALCON (v0.5) assemblers [35]. For FALCON, we adopted parameter sweeping and the assembly with the largest N50 was retained. For the other assemblers, default parameters were used, except the expected haploid genome size was set to values estimated by flow cytometry as well as k-mer analysis (Additional file 3). Quiver from SMRT Analysis v2.3.0 was run two times to polish base calling of assembled contigs [74].

Optical map construction

Long-range scaffolding of the assembly contigs with optical mapping was achieved using the Irys optical mapping platform (BioNano Genomics). HMW DNA was isolated from 3-week-old leaf tissue of in vitro grown 60444 and TME3 cassava plants, embedded in thin agarose plugs according to the IrysPrep Kit and the plant tissue DNA isolation protocol (BioNano Genomics). DNA molecules were labeled using the *NT.BspQI* DNA-nicking enzyme by incorporation of fluorescent-dUTP nucleotides according to the IrysPrep nick-and-repair protocol (BioNano Genomics). DNA samples were aliquoted and quantitated using the Qubit Fluorimeter run in broad-range mode. The final samples were then

loaded onto the IrysChips, linearized and visualized by the BioNano Irys molecule imaging instrument. Molecules > 150 kb were assembled de novo using the pairwise assembler provided by the IrysView software package (BioNano Genomics) with p value threshold of 10^{-9}.

Three-dimensional genome-wide chromatin capture sequencing

Freshly harvested leaves of in vitro grown cassava 60444 and TME3 plants were vacuum infiltrated in nuclei isolation buffer (NIB) supplemented with 2% formaldehyde. Protein crosslinking was stopped by adding glycine and applying an additional vacuum infiltration step. Leaf tissue was snap-frozen using liquid nitrogen and ground into a fine powder, re-suspend in NIB, and purified by spin-downs as described earlier [75]. Nuclei were digested with 400 units of *HindIII* as described in [75]. Digested chromatin was labeled using a fill-in reaction with 60 units of Klenow polymerase and biotin-14-dCTP. The exonuclease activity of T4 DNA polymerase was used to remove biotin-14-dCTP from non-ligated DNA ends. Proteinase K was added to reverse the formaldehyde crosslinking, and DNA was purified following phenol-chloroform extraction [75]. The Hi-C samples were quality assessed by PCR amplification of a 3C template and evaluated according to [75] (Additional file 1: Figure S3). Quality control passed Hi-C samples were purified following a phenol-chloroform extraction protocol [75] and mechanically sheared to fragment sizes of 300 bp using a Covaris S2 sonicator. Hi-C library fragments were blunt-ended using the End Repair Mix from Illumina and finally purified using AMPure beads according to the standard AMPure protocol. The biotinylated Hi-C samples were enriched through biotin-streptavidin-mediated pull-down and adenylated using Illumina's A-tailing mix. Illumina paired-end sequencing adaptors were ligated to the Hi-C fragments, and a PCR amplification of the Hi-C library was carried on as suggested earlier [75]. Finally, PCR products were purified using AMPure beads following the standard AMPure protocol and quantified using a Qubit device. Samples were sequenced using the Illumina HiSeq 2500 instrument. This produced 385 million pairs of 150-bp reads for 60444 and 391 million reads for TME3 (Additional file 2: Tables S13 and S14). Genome scaffolding was performed with Dovetail Genomics' HiRise scaffolding software.

Assembly accuracy estimation, repeat identification, and gene annotation

Publicly available WGS Illumina paired-end reads [76] were trimmed and quality filtered using Trimmomatic [77] and mapped to the draft assembly using BWA ALN (v0.7.12) [78] with default parameters. WGS read-mapping files were sorted using SAMtools SORT [79] statistics and called using QUALIMAP BAMQC [80]. Identification allelic sequences in all drafts was performed using Purge Haplotigs [39] (Additional file 1: Figure S16). To assess the assembly completeness, the set of reference CDSs (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta) was aligned to each of the assembled draft genome using GMAP [43] with option "-no fails" and "min-identity 0.5." Results were further filtered for alignments covering > 99% of query sequence using a custom script.

Repeat families found in the draft genome assemblies of 60444 and TME3 were first independently discovered de novo and structure classified using the software package REPEATMODELER ver. 1.0.9 and REPEATMASKER ver. 4.0.7 (https://www.repeatmasker.org). To screen for large tandem repeats, we used the software package RefAligner from Bionano with the option "-simpleRepeat -simpeRepeatTolerance 0.1 -simpleRepeatMinEle 3."

To annotate the gene space, we performed iterative MAKER analysis. In the initiated analysis, the gene prediction tool AUGUSTUS [81] was trained with reference gene models. The predicted gene models were combined with alignment base evidence, including all ESTs from cassava found on NCBI (https://www.ncbi.nlm.nih.gov/nucest/?term=cassava%20ESTs), Iso-Seq data, and UniProt protein sequences. The initiated set of MAKER gene models were used to train gene predictor SNAP, which was added in the second round of MAKER analysis, together with gene predictor GeneMark trained using Iso-Seq data. Putative gene functions of the final set of gene models were characterized by performing a BLAST search of the protein sequences against the Uniprot database (ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/). PFAM domains, InterProScanID, and Gene Ontology annotation were obtained by running interproscan [82]. To annotate non-protein-coding genes, the tools tRNAscan-SE [83] and Infernal [84] were used together with the Rfam version 13.0 database.

Allele-specific expression analysis and promoter region comparison

Newly generated RNA-seq datasets were derived from three key developmental stages of cassava 60444: early stage plant with fibrous root (FR) and leaf, middle stage plant with leaf, FR and intermediate root (IR), and late stage plant with leaf, FR, IR. RNA-seq libraries were sequenced using Illumina HiSeq2000 in paired-end 2 × 100 nucleotides mode. We aligned the RNA-seq reads using STAR [85] and retained the unique alignments. Reads were counted using SAMtools and custom made scripts [79].

Promoter regions were characterized for genes with two alleles and fpkm expression ratio > 0. Sequences 2 kb upstream of the start codon were defined as promoter. A pairwise alignment was generated for each allele pair using

the MUSCLE pairwise alignment tool [86]. Alignments were analyzed using 100-bp bins, and a similarity ratio was calculated using a custom script and visualized using the INCHLIB cluster and heatmap tools [87].

Genome-wide comparison and structural variation detection

To compare the 60444 and TME3 assemblies on a genome-wide scale, we used the optical maps of the two cassava cultivars to detect structural variations (SVs) using the RunBNG software [88]. We used the maps from 60444 as the reference and TME3 as query. RunBNG acts as a wrapper and essentially uses the BioNanos' RefAligner for generating the alignments. Alignments were then screened using the script "SVdetect" to detect the intergenomic SVs and to calculate insertion and deletion sizes [73]. Synteny was analyzed using the CoGe platform (https://genomevolution.org/). Syntenic regions between 60444 and TME3 were identified using CoGe SynMap and SynFind. The resulting table contains all genes in TME3 and the syntenic genes that were detected in 60444. We then defined a microsynteny score for every gene j in TME3. In a window of m genes surrounding gene j, stretching maximally n genes upstream and maximally n genes downstream on the same scaffold, we calculated for every m gene the longest syntenic gene sequence where all genes are conserved syntenic in the same or antisense direction in 60444. For $n = 5$, the maximal value per gene is thus 11 if gene j has both 5 genes up and 5 genes downstream and all 11 genes can be found in the same or antisense order in 60444. We then summed all scores of the genes in the window and divided by the square of the number of genes. Thus, in a window of 11 genes ABCDEFGHIJKTME3 where ABCDETME3 can be found in 60444 on Scaffold 1 and FGHIJKTME3 on Scaffold 2, the score is $5 \times 5 + 6 \times 6/11^2 = 0.504$. The same scoring results of a gene duplication in one genome but not the other.

The QTL *CMD2* on 60444 and TME3 has been identified using BLAST alignments of markers from the composite genetic map of cassava [33] and screened for markers from scaffold5214 and scaffold06906. Scaffold5214 has been reported by Rabbi and colleagues [16] to be closely linked to *CMD2*, and Scaffold6906 has been revealed in an association study [22]. Best BLAST hits were filtered and plotted using custom R-scripts. To identify the *CMD2* region of the AM560 genome, we used BLAST searches using a subset of the genetic markers: (1) Rabbi et al. [16] marker S5214_780931, (2) Wolfe et al. [22, 42] (only those with a p value < 10−50) S8_5645072, S8_5801843, S8_5801851, S8_6106055, S8_6218789, S8_6222418, S8_7325190, S8_7325312, S8_7325397, S8_7717243, S8_7717285, S8_7762525, S8_7762556, S8_7790078, S8_7790133. The markers represent SNPs; thus, a 81-bp region (40 bp either side of the disease resistance associated SNP) was used for each BLAST search. For each SNP marker, we performed a manual investigation and a single hit was identified on chromosome 12 and the *CMD2* locus was defined 100,000 bp either side of these BLAST hits.

Gene family analysis

To investigate gene family expansion specific in the 60444 or TME3 genomes, we used OrthoMCL clustering of all gene models present in our assemblies, the assembly of AM 560, the assembly of *Ricinus communis* as a close relative of cassava, and *Arabidopsis* as an outgroup [56, 57]. Only the longest protein sequence was selected, and datasets were filtered for internal stop codons. Pairwise sequence similarities between all input protein sequences were calculated using BLASTP [89] with an e value cutoff of 10^{-5}. Clustering of the resulting matrix was used to define the orthology cluster with an inflation value set to 1.5. Over- and underrepresentation of Gene Ontology (GO) terms between the three cassava genomic compartments were calculated with a hypergeometric test using the functions GOstats and GSEABase from the Bioconductor R package [90]. The REVIGO [91] package was used to remove redundant and similar terms from long Gene Ontology lists by semantic clustering and to visualize the enrichment results. To define local duplicated genes, OrthoMCL clustering was used. Local duplicated genes were reported when one orthologous neighboring gene was encoded on the same scaffold with a maximum distance of 100 kb and a 10 gene interval.

Enzyme prediction and pathway prediction was performed as published earlier [57].

Additional files

Additional file 1: Figure S1. Summary of data generated for genome construction. **Figure S2.** Genome size estimation for the two cassava genotypes using flow cell cytometry. **Figure S3.** Quality controls for the Hi-C libraries constructions. **Figure S4.** Pseudo-molecule validation using the 22,403 genetic markers from the cassava composite genetic map and the 18 pseudo-chromosomes of the cassava composite genetic map. **Figure S5.** Recombination rates for the cassava chromosomal pseudo-molecules. **Figure S6.** Plot of the length of the 18 chromosomes of AM560 compared to the combined length of the sequences that can be associated with the respective chromosomes in 60444 and TME3. **Figure S7.** Occurrence of genetic markers identified on TME3 and 60444. **Figure S8.** Genetic distance to physical distance plot of TME3 Scaffold 7, representing Chromosome 12 in AM560. **Figure S9.** Example of a misassembly identification using chromosome conformation capture read pairs. **Figure S10.** Summary of full-length transcriptome sequencing for high-quality gene-space annotation. **Figure S11.** GO enrichment analysis for the genes specific to the AM560 genome. **Figure S12.** GO enrichment analysis for the genes specific to the 60444 and TME3 genome. **Figure S13.** Squalene monooxygenase activity pathway and the corresponding gene models found in 60444, TME3 and AM560. **Figure S14.** Syntenic dotplot. **Figure S15.** Syntenic relation of the long arm of chromosome 12 between the AM560 v6.1 genome and equivalent

scaffolds of the TME3 or 60444 genomes. **Figure S16.** Read coverage histograms of TME3 and 60444 assemblies. (DOC 11613 kb)

Additional file 2: Table S1. Assembly statistics of representative genome drafts from the three different assemblers. **Table S2.** Assembly accuracy evaluation using publicly available Illumina paired-end reads. **Table S3.** Optical map assembly using the IrysView software provided by BioNano and using option 'optArguments_human'. **Table S4.** Structural variations based on optical maps of two cassava lines. **Table S8.** PacBio Iso-seq full length-transcriptome sequence classification. **Table S9.** Structural annotation of transposable elements in 60444 and TME3. **Table S10.** Non-coding RNA detected in the two cassava genomes. **Table S11.** BUSCO analysis of genome assemblies for 60444 and TME3. **Table S12.** Scaffolds representing the 18 pseudochromosomes of the cassava de novo genomes. **Table S13.** Hi-C library sequencing and read quality. **Table S14.** Hi-C read pair evaluation using HiCUP analysis pipeline (v0.5.8). (DOC 189 kb)

Additional file 3: Supplementary Materials and Methods. (DOC 177 kb)

Additional file 4: Table S5. Duplicated genetic markers in TME3 genome assemblies. **Table S6.** Duplicated allelic sequences (haplotigs) lifted up into TME3 pseudochromosome scaffolds by Dovetail using Hi-C data. **Table S7.** Duplicated allelic sequences (haplotigs) lifted up into 60444 pseudochromosome scaffolds by Dovetail using Hi-C data. (XLS 1479 kb)

Abbreviations

CaMV: Cauliflower mosaic virus; CDS: Coding DNA sequence; CM: Centimorgan; CMD: Cassava mosaic disease; FGCZ: Functional Genomic Center Zurich; FPKM: Fragments per kilobase of sequence per million mapped reads; FR: Fibrous root; GO: Gene Ontology; HMW: High molecular weight; INDELs: Insertions and deletions; IR: Intermediate root; LINE: Long interspersed element; LTR: Long terminal repeat; MYMV: Mungbean yellow mosaic virus; NCBI: The National Center for Biotechnology Information; NIB: Nucleus isolation buffer; PDI: Protein disulfide-isomerase; PE: Paired-end; PGDB: Plant genome database Japan; PTGS: Posttranscriptional gene silencing; RDR6: RNA-directed RNA polymerase 6; RE: Repetitive DNA element; R-genes: Resistance genes; SGS3: Suppressor of Gene Silencing 3; SINE: Short interspersed element; SMRT: Single-molecule, real-time sequencing; SRA: Short Read Archive; sRNA: Short RNA; SV: Structural variation; tDNA: DNA transposon; TEs: Transposable elements; TME: Tropical *Manihot esculenta*; TYLCV: Tomato yellow leaf curl geminivirus

Acknowledgements

We acknowledge Anna Bratus for the technical support at the Functional Genomic Center Zurich (FGCZ). We thank John Baeten (BioNano Genomics), Bo Xue, and Peifen Zhang (creation of PGDBs) for the technical assistance and the Dovetail Genomics team for assembling pseudochromosomes using Hi-C data. We thank Irene Zurkichen for the greenhouse support. We especially thank Rebecca Bart and her team at the Donald Danforth Plant Science Center, St. Louis, for sharing their information on the cassava TME7 genome, and Dario Copetti and Alexis Sarazin for the helpful scientific discussions.

Authors' contributions

JK, WQ and WG conceived the genome sequencing strategies for cassava cultivars 60444 and TME3, and HV initiated the *CMD2* research. JK prepared DNA samples for PacBio SMRT sequencing, AP performed PacBio library preparation and sequencing, JK and LP generated the BioNano optical genome maps, JK and SG generated Hi-C libraries with advice from UG, JK generated the Iso-Seq libraries, and AP performed Isoform Sequel sequencing. MK generated RNA-seq libraries. RSI performed genome size measurements. WQ performed the PacBio genome assemblies, BioNano scaffolding, gene space annotation and resolved haplotypes in all genome drafts. PRB analyzed annotations. MHH, JK and WG analyzed allele-specific expression data. JK performed transposable element analysis. JK and PS analyzed GO-annotation and synteny. PS performed enzyme and pathway prediction and their analysis. MHH and WQ contributed to the data release. JK, WQ, PS and WG analyzed the data and wrote the paper. All authors read and approved the final manuscript.

Funding

The work was supported by the Bill & Melinda Gates Foundation. J.K. was supported in part by a Swiss National Science Foundation grant to H.V.

Availability of data and materials

The cassava TME 3 and 60444 PacBio raw reads have been deposited at NCBI Short Read Archive (SRA) under BioProject number PRJEB27129 [92]. Genome assemblies and optical maps have been deposited at NCBI under BioProject number PRJNA508471 [93]. All other data are available from the corresponding authors upon reasonable requests. Public Illumina data sets SRX1393211 [94] and SRX526747 [76] were downloaded from NCBI SRA.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interest.

Author details

[1]Department of Biology, Institute of Molecular Plant Biology, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland. [2]Functional Genomics Center Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. [3]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland. [4]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. [5]AgroBioChem Department, University of Liège, Passage des Déportés 2, Gembloux, Belgium. [6]Advanced Plant Biotechnology Center, National Chung Hsing University, 145 Xingda Road, Taichung 40227, Taiwan.

Received: 27 June 2019 Accepted: 30 August 2019
Published online: 18 September 2019

References

1. Parmar A, Sturm B, Hensel O. Crops that feed the world : production and improvement of cassava for food, feed , and industrial uses. Food Secur. 2017;9:907–27.
2. Balat M, Balat H. Recent trends in global production and utilization of bio-ethanol fuel. Appl Energy. 2009;86:2273–82.
3. Ceballos H, Iglesias CA, Pérez JC, Dixon AGO. Cassava breeding: opportunities and challenges. Plant Mol Biol. 2004;56:503–16.
4. Ceballos H, Pérez JC, Joaqui Barandica O, Lenis JI, Morante N, Calle F, et al. Cassava breeding I: the value of breeding value. Front Plant Sci. 2016;7:1–12.
5. Hanley-Bowdoin L, Bejarano ER, Robertson D, Mansoor S. Geminiviruses: masters at redirecting and reprogramming plant processes. Nat Rev Microbiol. 2013;11:777–88.
6. McCallum EJ, Anjanappa RB, Gruissem W. Tackling agriculturally relevant diseases in the staple crop cassava (Manihot esculenta). Curr Opin Plant Biol. 2017;38:50–8.
7. Legg JP, Thresh JM. Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment. Virus Res. 2000;71:135–49.
8. Legg JP, Owor B, Sseruwagi P, Ndunguru J. Cassava mosaic virus disease in east and central Africa: epidemiology and management of a regional pandemic. Adv Virus Res. 2006;67:355–418.
9. Rey C, Vanderschuren H. Cassava mosaic and brown streak diseases: current perspectives and beyond. Annu Rev Virol. 2017;4:429–52.
10. de Ronde D, Butterbach P, Kormelink R. Dominant resistance against plant viruses. Front Plant Sci. 2014;5:307.
11. Lapidot M, Karniel U, Gelbart D, Fogel D, Evenor D, Kutsher Y, et al. A novel route controlling begomovirus resistance by the messenger RNA surveillance factor pelota. PLoS Genet. 2015;11:1–19.
12. Verlaan MG, Hutton SF, Ibrahem RM, Kormelink R, Visser RGF, Scott JW, et al. The tomato yellow leaf curl virus resistance genes Ty-1 and Ty-3 are allelic and code for DFDGD-class RNA-dependent RNA polymerases. PLoS Genet. 2013;9:e1003399.
13. Yamaguchi H, Ohnishi J, Saito A, Ohyama A, Nunome T, Miyatake K, et al. An NB-LRR gene, TYNBS1, is responsible for resistance mediated by the Ty-2 Begomovirus resistance locus of tomato. Theor Appl Genet. 2018;131:1345–62.

14. Okogbenin E, Egesi CN, Olasanmi B, Ogundapo O, Kahya S, Hurtado P, et al. Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. Crop Sci. 2012;52:2576–86.
15. Akano O, Dixon O, Barrera E, Fregene M. Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. Tag Theor Appl Genet Theor Und Angew Genet. 2002;105:521–5.
16. Rabbi IY, Hamblin MT, Kumar PL, Gedil M a., Ikpan AS, Jannink JL, et al. High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. Virus Res 2014;186:87–96.
17. Fondong VN. The search for resistance to cassava mosaic geminiviruses: how much we have accomplished, and what lies ahead. Front Plant Sci. 2017;8:1–19.
18. Beyene G, Chauhan RD, Wagaba H, Moll T, Alicai T, Miano D, et al. Loss of CMD2-mediated resistance to cassava mosaic disease in plants regenerated through somatic embryogenesis. Mol Plant Pathol. 2016;17:1095–110.
19. Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat Biotechnol. 2016;34:562–70.
20. Wang W, Feng B, Xiao J, Xia Z, Zhou X, Li P, et al. Cassava genome from a wild ancestor to cultivated varieties. Nat Commun. 2014;5:5110.
21. Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, et al. The cassava genome: current progress, future directions. Trop Plant Biol. 2012;5:88–94.
22. Wolfe MD, Rabbi IY, Egesi C, Hamblin M, Kawuki R, Kulakow P, et al. Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. Plant Genome. 2016;9:1–13.
23. Kayondo SI, Del Carpio DP, Lozano R, Ozimati A, Wolfe M, Baguma Y, et al. Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta. Sci Rep. 2018;8:1–11.
24. Masumba EA, Kapinga F, Mkamilo G, Salum K, Kulembeka H, Rounsley S, et al. QTL associated with resistance to cassava brown streak and cassava mosaic diseases in a bi-parental cross of two Tanzanian farmer varieties, Namikonga and Albert. Theor Appl Genet. 2017;130:2069–90.
25. Wilson MC, Mutka AM, Hummel AW, Berry J, Chauhan RD, Vijayaraghavan A, et al. Gene expression atlas for the food security crop cassava. New Phytol. 2017;213:1632–41.
26. Amuge T, Berger DK, Katari MS, Myburg AA, Goldman SL, Ferguson ME. A time series transcriptome analysis of cassava (Manihot esculenta Crantz) varieties challenged with Ugandan cassava brown streak virus. Sci Rep. 2017;7:1–21.
27. Anjanappa RB, Mehta D, Okoniewski MJ, Szabelska-Beręsewicz A, Gruissem W, Vanderschuren H. Molecular insights into cassava brown streak virus susceptibility and resistance by profiling of the early host response. Mol Plant Pathol. 2018;19:476–89.
28. Wang H, Beyene G, Zhai J, Feng S, Fahlgren N, Taylor NJ, et al. CG gene body DNA methylation changes and evolution of duplicated genes in cassava. Proc Natl Acad Sci. 2015;112:13729–34.
29. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329–46.
30. Staňková H, Hastie AR, Chan S, Vrána J, Tulpová Z, Kubaláková M, et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol J. 2016;14:1523–31.
31. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50.
32. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
33. (ICGMC) ICGMC. High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from ten populations. G3. 2015;5:133–44.
34. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
35. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13:1050.
36. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33:623–30.
37. Small KS, Brudno M, Hill MM, Sidow A. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. Genome Biol. 2007;8:R41.
38. Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, et al. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. Genome Res. 2005;15:1127–35.
39. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460.
40. Rival L, McKey D. Domestication and diversity in Manioc (*Manihot esculenta* Crantz ssp. *esculenta*, Euphorbiaceae). Curr Anthropol. 2008;49:1119–28.
41. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol. 2012;30:771–6.
42. Wolfe MD, Kulakow P, Rabbi IY, Jannink J-L. Marker-based estimates reveal significant non-additive effects in clonally propagated cassava (*Manihot esculenta*): implications for the prediction of total genetic value and the selection of varieties. G3. 2016;6:3497–506.
43. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.
44. Rojas MC, Pérez JC, Ceballos H, Baena D, Morante N, Calle F. Analysis of inbreeding depression in eight S1 cassava families. Crop Sci. 2009;49:543–8.
45. Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat Genet. 2017;49:1–7.
46. Sémon M, Wolfe KH. Consequences of genome duplication. Curr Opin Genet Dev. 2007;17:505–12.
47. Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikit S, Song C, et al. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat Commun. 2017;8:14953.
48. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, et al. Origin and evolution of the octoploid strawberry genome. Nat Genet. 2019;51:541–7.
49. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8:272–85.
50. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature. 2015;527:508–11.
51. Daccord N, Celton J-M, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet. 2017;49:1099–106.
52. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544:1–43.
53. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. *The Sorghum bicolor* genome and the diversification of grasses. Nature. 2009;457:551–6.
54. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of *Chenopodium quinoa*. Nature. 2017;542:1–6.
55. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
56. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
57. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol. 2017;173:2041–59.
58. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:258D–261.
59. Chen G, Pan H, Xie W, Wang S, Wu Q, Fang Y, et al. Virus infection of a weed increases vector attraction to and vector fitness on the weed. Sci Rep. 2013;3:1-6.
60. Kong L, Li S, Liao Q, Zhang Y, Sun R, Zhu X, et al. Oleanolic acid and ursolic acid: novel hepatitis C virus antivirals that inhibit NS5B activity. Antivir Res. 2013;98:44–53.
61. Ashour A, El-Sharkawy S, Amer M, Abdel Bar F, Katakura Y, Miyamoto T, et al. Rational design and synthesis of topoisomerase i and II inhibitors based on oleanolic acid moiety for new anti-cancer drugs. Bioorganic Med Chem. 2014;22:211–20.

62. Patil BL, Fauquet CM. Cassava mosaic geminiviruses: actual knowledge and perspectives. Mol Plant Pathol. 2009;10:685–701.
63. Pant V, Gupta D, Choudhury NR, Malathi VG, Varma A, Mukherjee SK. Molecular characterization of the Rep protein of the blackgram isolate of Indian mungbean yellow mosaic virus. J Gen Virol. 2001;82:2559–67.
64. Bull SE, Briddon RW, Sserubombwe WS, Ngugi K, Markham PG, Stanley J, et al. Genetic diversity and phylogeography of cassava mosaic viruses in Kenya. J Gen Virol. 2006;87:3053–65.
65. Almagro L, Gómez Ros LV, Belchi-Navarro S, Bru R, Ros Barceló A, Pedreño MA. Class III peroxidases in plant defence reactions. J Exp Bot. 2009;60:377–90.
66. Dieng H, Satho T, Hassan AA, Aziz AT, Morales RE, Hamid SA, et al. Peroxidase activity after viral infection and whitefly infestation in juvenile and mature leaves of *Solanum lycopersicum*. J Phytopathol. 2011;159:707–12.
67. Yang P, Lüpken T, Habekuss A, Hensel G, Steuernagel B, Kilian B, et al. PROTEIN DISULFIDE ISOMERASE LIKE 5-1 is a susceptibility factor to plant viruses. Proc Natl Acad Sci U S A. 2014;111:2104–9.
68. Mourrain P, Béclin C, Elmayan T, Feuerbach F, Godon C, Morel J-B, et al. Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. Cell. 2000;101:533–42.
69. Maine EM. A conserved mechanism for post-transcriptional gene silencing? Genome Biol. 2000;1:1018.1–4.
70. Glick E, Zrachya A, Levy Y, Mett A, Gidoni D, Belausov E, et al. Interaction with host SGS3 is required for suppression of RNA silencing by tomato yellow leaf curl virus V2 protein. Proc Natl Acad Sci U S A. 2009;106:4571.
71. Horvath P, Barrangou R, Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Source Sci New Ser. 2010;327:167–70.
72. Doyle JJ, Doyle JL. A rapid total DNA preparation procedure for fresh plant tissue. Focus (Madison). 1990;12:13–5.
73. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015;12:780–6.
74. Chin C, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563–9.
75. Grob S, Grossniklaus U. Chromatin conformation capture-based analysis of nuclear architecture. In: Kovalchuk I, editor. Plant epigenetics: methods and protocols. Boston: Springer US; 2017. p. 15–32.
76. DOE-Joint Genome Institute. cassava WGS sequencing.2015. https://www.ncbi.nlm.nih.gov/sra/?term=SRX526747
77. Bolger AM, Lohse M, Usadel B. Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
78. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
79. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
80. Okonechnikov K, Conesa A, Garcı F. Genome analysis Qualimap 2 : advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2015;2016(32):292–4.
81. Stanke M, Diekhans M, Baertsch R, Haussler D. Sequence analysis using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
82. Conesa A, Götz S, García-gómez JM, Terol J, Talón M, Genómica D, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
83. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016;44:54–7.
84. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
85. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
86. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
87. Škuta C, Bartu°něk P, Svozil D. InCHlib - interactive cluster heatmap for web applications. J Cheminform. 2014;44:1–9.
88. Yuan Y, Bayer PE, Lee H, Edwards D. Sequence analysis runBNG : a software package for BioNano genomic analysis on the command line. Bioinformatics. 2017:1–3.
89. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
90. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
91. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6.
92. Functional Genomics. Center Zurich. Cassava genomes assembled with single-molecule long reads, optical and Hi-C maps reveal narrow genetic diversity and mono-allelic expression. 2018. https://www.ncbi.nlm.nih.gov/bioproject/PRJEB27129
93. Kuon JE, Qi W, and Gruissem W. Cassava genomes assembled with single-molecule long reads, optical and Hi-C maps. 2018. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA508471
94. DOE-Joint Genome Institute. cassava WGS sequencing.2015. https://www.ncbi.nlm.nih.gov/sra/?term=SRX1393211

Publisher's Note